DECISION RULES BASED ON HYPOTHESIS TESTS AND EFFECT SIZES FOR LOGISTIC REGRESSION DIFFERENTIAL ITEM FUNCTIONING

by

Adam Gesicki

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate and Postdoctoral Studies

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA (Vancouver)

September 2015

© Adam Gesicki, 2015
**Abstract**

Logistic Regression (LR) has been a technique used for the detection of items exhibiting differential item functioning (DIF). When it was introduced in 1990, the LR was conceptualized as strictly a test of statistical significance. This led to the over-identification of items as DIF, generally not exhibiting practically (psychometrically) significant differences. The use of blended decision rules – where effect sizes are used in addition to statistical significance in the decision-making process – was proposed to address this issue. Previous work in the literature attempted to align a decision rule grounded in the Mantel-Haenszel (M-H) technique to LR. However, this work is unable to replicate previously recommended cut-offs, through the use of the same methodology on a different data set. It is possible that cut-off values may be dataset specific, which also opens the question of whether universal cut-off values for effect sizes for DIF are a realistic expectation.
Preface

This dissertation is the original independent work by the author, Adam Gesicki.
Table of Contents

Abstract .......................................................................................................................... ii
Preface ........................................................................................................................... iii
Table of Contents .......................................................................................................... iv
List of Tables ................................................................................................................... v
List of Figures .................................................................................................................. vi
List of Abbreviations ..................................................................................................... vii
Introduction ................................................................................................................... 1
Background Literature .................................................................................................... 5
   The Mantel-Haenszel (M-H) Technique ................................................................. 5
   The Logistic Regression (LR) Technique .................................................................. 7
   Advantages of the LR Method Over the M-H ....................................................... 9
DIF: Defined by Statistical Significance, Effect Size, or Both? ................................. 11
Development of Recommended Cut-Offs for the $R^2$ ............................................. 14
Applications of Decision Rules in Simulation Studies .............................................. 17
Statement of the Research Questions ......................................................................... 19
Methods ....................................................................................................................... 22
   Data Source .............................................................................................................. 22
   Question 1: Translating M-H Cut-Offs to LR ....................................................... 24
   Question 2: Comparability of Results .................................................................... 25
Results ......................................................................................................................... 26
   Results for Question 1: Translating M-H Cut-Offs to LR..................................... 26
   Results for Question 2: Comparability of Results ................................................. 36
Concluding Remarks ..................................................................................................... 42
   Implications for Practice ....................................................................................... 44
   Other Directions for Research .............................................................................. 46
References ...................................................................................................................... 47
List of Tables

Table 1. Sample size under various conditions for the four forms under consideration in this work. ................................................................. 23

Table 2. A table of exact agreement (off-diagonal) and proportions (diagonal) of various decision rules applied to all types of DIF under consideration................................. 37

Table 3. A table of exact agreement (off-diagonal) and proportions (diagonal) of various decision rules applied to gender DIF data only................................................................. 38

Table 4. Classification contingency matrix for DIF items using decision rules from Holland & Thayer (1988) for M-H vs. those calculated in this work, all numbers as proportions of total item mix for all data......................................................................................................................... 40

Table 5. Summary of differences in cut-offs on the $\Delta R^2$ metric between various sources used to-date in the literature. ................................................................................................. 43
List of Figures

Figure 1. A contingency matrix for use in cases of null-hypothesis statistical testing. .................. 12

Figure 2. A contingency matrix for use in cases of a blended decision rule. .............................. 13

Figure 3. Density plot of summed score by gender, in both listening forms (left) and reading forms (right). ................................................................. 27

Figure 4. Density plot of summed score by sociocultural background, in both listening forms (left) and reading forms (right). ............................................................. 28

Figure 5. SIBTEST $|\hat{\beta}|$ on x-axis with LR-based $\Delta R^2$ between models 1 and 2 on the y-axis.... 29

Figure 6. Rescaling of Figure 5 to match the presentation of Figure 1 from Jodoin and Gierl (2001, p. 336). ........................................................................................................... 30

Figure 7. Relationship between deltaMH and the logOdds of Group in model 2, placed on the same metric scale, with superimposed linear regressions. Points representing gender DIF are clustered towards the bottom-right of the graph and are represented with circles, with socioculturally data clustered to the top-left of gender and represented with triangles ................................................................................................................................... 32

Figure 8. A graph demonstrating the median results of the increasing window around the target logOdds of 0.42553. The dotted horizontal line demonstrates the model-based estimate, discussed later in text ................................................................................................................................. 34

Figure 9. The relationship between the absolute value of the logOdds of group in model 2 as opposed to $\Delta R^2$ between models 1 and 2. The loess curves are plotted using solid lines; matching modelled relationships (power) in dotted lines ................................................................................................................................. 35
Figure 10. A graph comparing the distribution of $\Delta R^2$ s between Models 2 and 3. The top graph shows the distribution of $\Delta R^2$ for items classified equal / lower using decision rules suggested here, when compared to those suggested by Holland & Thayer. The bottom graph shows the distribution of $\Delta R^2$ for items classified higher (containing more DIF) using the new decision rule as opposed to their Holland & Thayer classification. ....... 41
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIF</td>
<td>Differential Item Functioning</td>
</tr>
<tr>
<td>LR</td>
<td>Logistic Regression</td>
</tr>
<tr>
<td>M-H</td>
<td>Mantel-Haenszel</td>
</tr>
<tr>
<td>SIBTEST</td>
<td>Simultaneous Item Bias Test</td>
</tr>
</tbody>
</table>
Introduction

Differential item functioning, or DIF, is an important consideration for the test developer, as highlighted by the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 2014; referenced as the *Standards* for the remainder of this work). Camilli (1993) clearly defined DIF, stating:

> An item is said to “function differently” for two or more groups if the probability of a correct answer to a test item is associated with group membership for examinees of comparable ability. Statistical indices of DIF are designed to identify such test items. If the degree of DIF is determined to be practically significant for an item and the DIF can be attributed plausibly to a feature of the item that is irrelevant to the test construct, the presence of this item on the test biases the ability estimates of some individuals. This compound condition, when satisfied, indicates item bias. (pp. 397-8).

What is most apparent from this definition is that DIF is simply a statistical index – an indicator – that signals that an item may be biased. It is only the follow-up of content experts that can determine whether or not the statistical flag is as a result of a biased item, and what features of the item may be responsible for such a difference between test-taker groups. As well, it is important to highlight from the definition that any statistic used for DIF detection must only compare test takers of similar ability. If test taker abilities are not taken into account in the calculation, one cannot isolate whether or not differences in an item’s performance are due to features of the item or a difference in the underlying ability of test takers. Lastly, DIF, as the name suggests, specifically targets analysis at the *item* level. Further extensions of DIF to the testlet or test form level exist, but are not directly discussed here.
In discussing definitions surrounding DIF, it is also important to differentiate between item bias and item impact. Impact is defined when a detected “difference in group performance” was “caused by valid skill group differences” (Ackerman, 1992, p. 6). In other words, where there “were true differences between the groups in the underlying ability of interest being measured by the item” (Zumbo, 2007, p. 224). Naturally, the use of item content panelists, together with the triangulation of other sources of data, would help test developers narrow which of the two conditions is present in a particular item. This judgement may – of course – not always be clear-cut. This process attempts to strengthen the validity of the testing process, by identifying items whose inclusion may ultimately lead to inappropriate inferences from a test-taker’s score. However, even discussions regarding bias and impact require a statistical technique that can support the process undertaken by content experts.

The variety of techniques used for the detection of DIF are neatly summarized in Millsap and Everson (1993). Other notable reviews include Zumbo’s (2007) three generations of DIF analyses, which reviewed where the field had been and provided a research agenda for the future.

This work specifically focuses on one of these methods, the logistic regression (LR) DIF technique. Introduced by Swaminathan and Rogers (1990), the LR DIF technique was presented as an alternative to the detection of DIF using the Mantel-Haenszel (M-H) test (Holland & Thayer, 1988). Regardless of the statistical technique under discussion, all forms require a decision rule to determine whether or not an item exhibits signs of DIF. As originally presented by Swaminathan and Rogers (1990), the decision rule for LR revolved solely around a statistical test of statistical significance. In this case, a statistically significant test result was to be interpreted as an item that had differential functioning. Eventually, the notion of a “blended” decision rule was introduced; that is, a rule where both the statistical significance of a test as well
as its effect size are considered as evidence in deciding whether an item is differentially functioning; mirroring a strategy used by M-H DIF since its inception (Zumbo, 2008; Holland & Thayer, 1988). Unlike the M-H, which was introduced with interpretable guidelines, the LR DIF technique does not have a consistently agreed-upon blended decision rule to determine whether or not an item should be classified as “DIF” vs. “non-DIF”. Despite this limitation, the LR DIF technique offers a number of qualities that M-H DIF does not offer, such as: the detection of non-uniform DIF (explained later), the capacity to treat the ability range as a continuous variable (as opposed to the discretization required under M-H DIF), and increased modelling flexibility. These benefits suggest that it would be worthwhile to consider the role of blended decision rules within LR DIF, and strive for harmonization of some of the suggestions made to-date in the literature.

As a result, the purpose of this work is to:

1. summarize existing approaches in the development of LR DIF blended decision rules;
2. attempt to replicate methodology suggested in the literature (Jodoin & Gierl, 2001) and recommend a different approach for a potential transfer of M-H DIF cut-offs into a scale used by LR DIF; and,
3. show the comparability between various decision rules.

The work will begin with a review of the background literature of the M-H and LR DIF techniques, as well as highlight previous recommendations made in the literature under the LR DIF framework. Next, another method will be described which will enable the existing and well-established M-H DIF blended decision rule to be transformed to a comparable rule within the LR DIF framework. The results section will provide preliminary results of the new cut-offs, especially highlighting the comparability of the different decision rules discussed in the
background literature using data derived from an actual testing programme. For the purposes of brevity throughout the remainder of the paper, “M-H DIF” and “LR DIF” will simply be referred to by “M-H” and “LR” methods respectively; their use as techniques for DIF detection are to be implied throughout the remainder of this work.
Background Literature

The Mantel-Haenszel (M-H) Technique

The Mantel-Haenszel (M-H) technique has been used extensively by many individuals in testing, especially due to its long history within the field. M-H began with the application of a statistical technique presented by Mantel and Haenszel (1959) into the testing realm by Holland and Thayer in 1988. The M-H categorizes the ability range into a number of levels \( k \); and within each, considers the number of the members within each of the test taker groupings R and F, who have answered the question correctly \( (N_{R1k}, N_{F1k}) \) and incorrectly \( (N_{R0k}, N_{F0k}) \). The estimate of the conditional odds ratio is calculated as:

\[
\hat{\alpha}_{MH} = \frac{\sum_k N_{R1k}N_{F0k}}{\sum_k N_{R0k}N_{F1k}}.
\]

The conditional odds ratio estimated by the M-H technique is usually subsequently transformed to a “delta” metric, denoted as \( \Delta_{\alpha_{MH}} \). This has the purpose of bringing the estimate of the conditional odds ratio onto the same scale used by ETS test developers to describe item

---

1 The symbols R (“reference”) and F (“focal”) groups are preserved in this section to respect the notation used in existing research, and could easily have been substituted for G1 (“group 1”) and G2 (“group 2”). As a philosophical note, the use of the terms “reference group” and “focal group” will be avoided in this paper. Traditionally, this terminology is associated with the “first generation” of DIF researchers, who were most interested in dominant and minority groups (respectively; Zumbo, 2007, p. 224). An acknowledgement is made that there existed many historical and legal reasons for such terminology to be used within the realm of DIF; however, this paper focuses on the use of terminology that does not intend to value any one group of test-takers over another.
difficulty, making it more interpretable for internal use (a metric with a normal distribution, with a mean of 13 and standard deviation of four; Holland & Thayer, 1985; Zwick, 2004). This is done through the following relationship:

$$\Delta_{\alpha_{MH}} = -2.35 \ln(\hat{\alpha}_{MH}).$$

This transformation also had the additional side-effect of linearizing the odds ratio (i.e., the difference between 0 and 0.5 represents the same difference as 0 to -0.5), as well as centering around zero (i.e., no-difference between the two groups being examined would be represented with a value of zero as opposed to one).

The statistical significance of the M-H is calculated with the help of the M-H chi-square test statistic:

$$\chi^2_{MH} = \left( \frac{\sum_k N_{Rik} - \sum k E(N_{Rik}) - 0.5}{\sum k Var(N_{Rik})} \right)^2,$$

Where:

$$E(N_{Rik}) = \frac{n_{Rik}m_{ik}}{N_k} \quad \text{and} \quad Var(N_{Rik}) = \frac{n_{Rik}n_{Fik}m_{ik}m_{0ik}}{N_k^2(N_k - 1)}.$$

The subscripts represent whether the number of test takers ($n$ or $m$) counted are those who:

- belong to a specific group membership (“R” vs. “F”), and
- those who got the item-in-question correct or incorrect (“1” vs. “0”, respectively).

The chi-square of the above test is compared against the critical value of the chi-square distribution, with one degree of freedom and an alpha level of 0.05.

The interpretation of the M-H test is generally done using the ETS-recommended decision rules, which when used, categorize each item into either A, B, or C. These three
categories represent increasing levels of detected DIF: A-level represents items to be considered free of DIF, B-level is to classify items with some DIF that are less favoured than A-level items, and C-level is used to describe items that should only be retained in the cases where it is “essential to meet test specifications” (Zwick & Ercikan, 1989, p. 59). More practically, items exhibiting C-level DIF are targeted for removal if there is no strong underlying argument for the item’s continued inclusion, whereas A-level items are interpreted to have either no, or minimal levels, of DIF (Zwick & Ercikan, 1989). The ETS decision rule behind the M-H is summarized as:

| To meet condition… | the statistical test must be… | $|\Delta_{\text{M-H}}|$ is… |
|--------------------|---------------------------------|----------------------|
| A                  | non-significant                 | or                   |
| B                  | statistically significant       | and                  |
| C                  | statistically significant       | and                  |

The Logistic Regression (LR) Technique

The LR method for the detection of DIF was first discussed by Swaminathan and Rogers (1990) as a potential alternative to the detection of DIF with M-H (see also Rogers & Swaminathan, 1993). Unlike the M-H test, the LR method does not have established guidelines for its interpretation beyond the tests of statistical significance. As a result, a discussion of the LR method, and its considerations will be detailed. The model proposed was the following binary logistic regression with logit link function:

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \theta + \beta_2 g + \beta_3 (\theta g)$$

with

- $P$ representing the probability of successfully answering the item under investigation for DIF,
• $\theta$ representing an individual’s ability,
• $\beta_0$ representing the regression intercept,
• $\beta_1$ the regression weight for the ability’s effect (shown by $\theta$) in log-odds,
• $\beta_2$ being the regression weight for group membership ($g$) in log-odds, and
• $\beta_3$ being the regression weight of the interaction effect between ability ($\theta$) and group membership ($g$) in log-odds.

The $\beta_2$ and $\beta_3$ terms are often termed the uniform and non-uniform (respectively) components of DIF of an item. Uniform DIF is used to define a difference in probability between the two test taker groups that is equidistant throughout the ability range, represented by the single ability variable. Non-uniform DIF represents where the difference of success on an item between two groups varies throughout the ability range, represented mathematically with an interaction term.

The ability term, $\theta$, can be estimated in a number of ways. The most common method is to use the summed total score on the measurement instrument (cf. Rogers & Swaminathan, 1990; Zumbo, 1999; Cuevas & Cervantes, 2012). It is relevant to note that, despite this common usage of regression for DIF, the assumption that the predictor variables are without error is, in fact, violated (Shear & Zumbo, 2013). More specifically, when the summed score is used as an estimator of ability, it is done so under a classical test theory framework, which suggests that the observed score is composed of both the true score and an uncorrelated random error (Shear & Zumbo, 2013). Reference to this warning is made here to serve as a reminder and acknowledgement; however, considering the current and active use of LR for the study of DIF, this work will continue in line with the usage found in the literature.
The original Swaminathan and Rogers publications (1990, 1993) suggested the use of a two degree-of-freedom $G^2$ test on the regression model presented above. A statistically significant value, at an alpha of 0.05, was proposed as the indicator for DIF. The initial 1990 publication did include a distributional study which verified that the LR test statistic for the two-degree of freedom test was indeed distributed according to the expected chi-square ($\chi^2$) distribution.

Zumbo (1999) further advanced the LR method by describing the original regression method detailed above as two consecutive, nested models. The three resulting models are: model 1: a model that only includes ability as predictor; model 2: a model that includes ability and group membership; and model 3: the final model with ability, group, and the interaction term between the two variables. The proposed advantage of consecutive one degree of freedom chi-square tests is that it allows a separate focus on uniform DIF (i.e., the comparison between model 1 and model 2) from non-uniform DIF (i.e., comparing model 2 to 3). This, in turn, provides the advantage that one does not lose a degree of freedom in doing the two-degree of freedom test on an item that only contains uniform DIF (Swaminathan & Rogers, 1990).

**Advantages of the LR Method Over the M-H**

Upon introduction of the LR method, early adopters often compared it to the dominant strategy of the time, the M-H, and the advantages it provided over the M-H. The largest difference noted by Swaminathan and Rogers (1990) was that the LR technique detects the presence of non-uniform DIF, as reflected by $\beta_3$, which is not expressly built into the Mantel-Haenszel statistic (Rogers & Swaminathan, 1993). This effect is not to be minimized as, in some cases, as much as 16% of items containing DIF have been suggested to contain non-uniform
differences (Maller, 2001). Moreover, the LR method does not require the arbitrary categorization of a continuous criterion (ability) variable, which is necessary in the creation of the contingency tables required in the M-H (Zumbo, 1999). The LR-based models can be expanded to be more complex, including multiple matching variables (such as multiple estimates of ability; French & Maller, 2007). Furthermore, conceptualizing DIF from an LR perspective allows a host of techniques from the logistic regression framework to be used, such as multinomial logistic regression (French & Maller, 2007) or logistic discriminant function analysis (Gómez-Benito, Hildago, & Zumbo, 2013). Lastly, LR is a statistical method that is freely available in most of today’s statistical computer programmes, making it a tool that can be easily implemented (Zumbo, 1999; French & Maller, 2007).

The original work also began the statistical comparison between the already existing M-H technique and the newly proposed LR technique. Rogers and Swaminathan’s (1993) study especially established that, in terms of power, LR did reasonably well when compared to the M-H for uniform DIF, and was better able to detect some non-uniform DIF (especially due to its expressed design in doing so). As well, the early studies noted differences that were expected: power was higher when the size of the DIF was greater, a longer test length increased the power of detection, and increased sample sizes increased power as well (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993). They went on to also identify that false-positive rates for LR were variable, ranging from 1% to 6% for LR, as opposed to the 1% that the M-H “consistently” performed.

Therefore, if the LR technique is to be a tool to be used for DIF detection – and especially as an alternative to M-H for the detection of DIF – the already established M-H blended decision rule must be transformed onto a metric that can be used by LR. Naturally,
understanding the nature of comparability between different tests is practically important. Identifying items incorrectly as DIF (either items that are differentially functioning but do not flag statistically, or vice-versa) can have serious implications. These could include the removal of items from an item bank at a high cost from an operational perspective, or the development of a different understanding of an underlying construct as items are removed incorrectly from a research perspective. High Type I error may be responsible for “seemingly innocuous items” which are statistically flagged for DIF (Russous & Stout, 1996, p. 215). Finally, the understanding of bias and the nature of DIF underlying items would be impacted, interfering with the call for deeper understanding of DIF made by Zumbo (2007). If the LR model is assumed to be a reasonable representation of DIF, then it stands to consider what decision rule(s) are to be used to convert the model’s results into decisions of practical significance.

**DIF: Defined by Statistical Significance, Effect Size, or Both?**

Effect sizes have been central to the DIF landscape since very early in the DIF literature. As discussed above, the M-H’s “A”, “B”, or “C” levels are based on the effect size measure of $|\Lambda_{\text{MH}}|$. This three-level terminology continues to be mirrored with other techniques; for example, A/B/C groupings also appear in Zumbo and Thomas (1997) and Jodoin and Gierl (2001). Often, in the literature, the drive to include a blended decision rule to the LR is fueled by a concern that chi-square tests of statistical significance are sensitive to large sample sizes, causing a situation where a large enough sample size may magnify a minute difference between two test taker groups (Gómez-Benito et al., 2013). This concern is echoed in Jodoin and Gierl (2001), where it is hoped that the inclusion of effect sizes may lower “Type I error rates” (p. 329); more specifically, “the identification of an item as displaying DIF when, in fact, it does not” (p. 330).
However, in some eyes, a trivial difference that is magnified with a large sample size is, in fact, considered to be an overpowered test statistic, rather than an issue of strict Type I error (Ellis, 2010; Lin, Lucas, Shmueli, 2013). In the context of DIF studies to-date, the literature has used a number of definitions of “Type I” and “Type II”; terms that have traditionally been reserved for null hypothesis significance testing. It is encouraged that further research in this area clearly distinguishes among the cells of a contingency matrix and differentiate the terms when the decision rule is based solely on hypothesis testing, as opposed to that of a blended decision rule.

<table>
<thead>
<tr>
<th>Test Outcome (Statistical Significance)</th>
<th>“Gold Standard” / Truth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>condition true</td>
</tr>
<tr>
<td><strong>DIF</strong></td>
<td>Correct Conclusion</td>
</tr>
<tr>
<td><strong>no DIF</strong></td>
<td>Type II Error</td>
</tr>
</tbody>
</table>

Figure 1. A contingency matrix for use in cases of null-hypothesis statistical testing.

When used in line with classical statistical testing (as described by Neyman-Pearson; Lehmann, 1993), the test under consideration is the test of statistical significance of the logistic regression, as conceptualized by Swaminathan and Rogers (1990), based on the sampled data. In this case, the matching gold standard needs to resemble a statistical one: one where the null hypothesis of no (exact) difference between the two groups under consideration exist in the population. Therefore, when discussions occur of sensitivity of the test to large sample size populations, the concern is really one of an over-powered test statistic (Lenth, 2001).
In the other case, when the *gold standard* is modified to represent DIF that is of practical (or psychometric) significance rather than an exact difference, the *test* used in this consideration should be clarified to mean a blended decision rule. In this case, the difference in the *gold standard* is reflected through the inclusion of an appropriate statistical measure: the effect size. To reflect the change in the nature of the test, it is recommended that the terms should appropriately be changed to those commonly used in statistical learning and not to view them as completely synonymous (James, Witten, Hastie, & Tibshirani, 2013; see Figure 2 above). Given the concerns noted in Jodoin and Gierl (2001), the *chi-square* test raises the issue of having too many false positives (FP). The use of the FP term is a better description of the actual concern when the *chi-square* test is discussed in the contest of DIF.

There exists a third option to consider for the *gold standard*, which has not been studied in the literature to-date: the agreement of a test’s results to that of a DIF panel of experts who agree in the test’s indication of a DIF item that requires removal from a testing instrument. This would still retain the language used in Figure 2, because a blended decision rule would still be in use.
Ultimately, as stated by Potenza and Dorans (1995): “To be used effectively, a DIF detection technique needs an interpretable measure of the amount of DIF” (p. 33). The same also appears in the Standards, where a reminder is drawn that analyses must detect “meaningful DIF” (AERA/APA/NCME, 2014, p. 89). This is in line with recommendations within the broader research literature to use effect sizes in enriching interpretations beyond statistical significance (Wilkinson & Task Force on Statistical Inference, 1999). With clarifications on the role of the blended decision rule in a contingency matrix in place, the following section will highlight attempts in the development of effect size cut-offs for the LR.

**Development of Recommended Cut-Offs for the $R^2$**

Although the M-H test has only one conceptualization of effect size in use, the LR method has multiple ways of portraying effect size. At the time of the original Swaminathan and Rogers studies (1990, 1993), the use of effect sizes was not discussed. In essence, the decision rule only consisted of a test of statistical significance. The notion of the use of effect sizes in LR was first introduced by Zumbo and Thomas (1997), which recommended that the effect size cut-offs on the $\Delta R^2$ scale should be 0.13 and 0.26, akin to the transition between A- and B- level, as well as B- to C-level DIF (respectively). These were informed by the effect size guidelines put forth by Cohen (1992) and represented a best-guess attempt at the creation of interpretable benchmarks for DIF without the presence of previously established guidelines for LR.

In an attempt to link the interpretation of the M-H and the LR techniques, Jodoin and Gierl (2001) converted the classification benchmarks established for the Simultaneous Item Bias (SIBTEST) method onto the $\Delta R^2$ effect size of LR; given that the SIBTEST cut-offs had originally been based on the M-H effect size measures (see Russous & Stout, 1996). To facilitate these conversions, Jodoin and Gierl (2001) calculated the relevant measures using data from:
• a 50-item grade six social studies achievement test, with the two groups defined by English-French translations of the same test, each group containing 2,220 test takers,
• a 50-item grade six mathematics achievement test, with groupings as above, and
• a 70-item grade 12 social studies high school exit exam, with gender DIF under consideration, but no number of test takers in each group specified.

Based on this data, Jodoin and Gierl (2001) noted that the relationships between the SIBTEST cut-offs and LR $\Delta R^2$ were similar regardless of test content or group split considered. This resulted in final recommendations of a $\Delta R^2$ under 0.035 for A, over or equal to 0.035 but under 0.070 for B, and greater than or equal to 0.070 for C. Again, following the tradition of the M-H, Jodoin and Gierl (2001) suggest that statistical significance must be met for B- and C-level DIF exclusively.

As can be seen by direct comparison, the cut-offs presented by Jodoin and Gierl (2001) represent quite a distinct reduction from those initially suggested by Zumbo and Thomas in 1997. The newly recommended cut-offs, as expected, greatly influenced the number of items detected as containing moderate DIF from 6.8% to 68.2% of DIF items (Jodoin & Gierl, 2001). From a power standpoint, this change strongly weighed in favour of the smaller effect size requirements, making the Jodoin and Gierl-suggested cut-offs most often used in recent published literature (cf. Hidalgo & López-Pina, 2004).

Jodoin and Gierl’s (2001) $\Delta R^2$ cut-offs were originally designed to be applied at each sequential regression model, testing uniform and non-uniform separately using the same $\Delta R^2$ cut-offs. The authors stated that, because their recommendations were derived from M-H (albeit via SIB), the cut-offs only are relevant to the transition between models one and two, which are
aimed to test for uniform DIF (the only type of DIF the M-H is designed to detect). The authors suggested that with no levels established by the psychometric community for non-uniform DIF, the same cut-offs should be applied to both non-uniform as to uniform DIF. Interestingly, the authors wrote earlier in the same study that “SIB was chosen as the basis of comparison because it is able to detect both uniform and non-uniform DIF, and it has superior statistical characteristics compared to MH in the detection of both uniform and non-uniform DIF” (p. 335; emphasis added). This is a stark contrast to the assertion that the cut-offs must be used separately for uniform as well as non-uniform DIF. Given that the M-H metric does report on conditional odds, it is also strange that the study authors do not directly compare the M-H odds with the odds reported by LR. This feature is explicitly stated in one of Holland and Thayer’s (1986) works: “The parameters, α and ln(α), are also called ‘partial association’ parameters because they are analogous to the partial correlations used with continuous data” (p. 9).

Lastly, it is worth discussing the unique contributions of Cuevas and Cervantes (2012), which slightly departed from the pattern of previous research in this field. This is primarily attributed to the study’s intent to analyze DIF in the environment of a Columbian testing programme. Unlike the studies detailed earlier, which generally detailed sample sizes under 1,000 test-takers per group, Cuevas and Cervantes (2012) considered a simulation population where the primary group had either 7,500, 8,500, 26,000, or 33,000 test-takers, with many additional ratios for the second group of test-takers. Previous research explicitly stated the use of the Nagelkerke $R^2$ (1991) for LR as well (see Zumbo, 1999; Jodoin & Gierl, 2001), whereas Cuevas and Cervantes (2012) selected the McFadden $R^2$ for their use (although they did not attempt to apply any of the existing cut-offs to their study).
Although the methodology for the Cuevas and Cervantes (2012) study departed from existing trends in the literature, there are a number of approaches discussed within the article that are useful to draw upon. Of note is the use of receiver operator characteristic (ROC) curves to determine optimality under each condition. This method generates a graph of the false positive rate on the $x$-axis, and true positive rate on the $y$-axis. Using this method allows the identification of a point on the curve closest to the top left of the graph; that point representing the ideal condition of no false positives and 100% true positives. Through the use of this method, it is interesting to see that the $\Delta R^2$ cut-offs suggested by the authors are different, depending on sample size. This is unusual considering the goal of exploring the effect sizes in the first place was to prevent sample size from influencing the results (see discussion earlier). The deeper implication within this reasoning is that sample size would change the maximum permissible amount of DIF in a particular item – something that is not operationally defendable, nor theoretically justifiable. This is also a strategy that is not supported with the current use of DIF results (such as the M-H). What is also striking about the cut-offs suggested in Cuevas and Cervantes (2012) is the large discrepancy when the results are compared to those put forward by Jodoin and Gierl (2001), let alone Zumbo and Thomas (1997). The suggested cut-offs for uniform DIF, in the small DIF condition (assuming a typographical error at the top of table 4 on p. 56), range from 0.000082 to 0.000363 (dependent on sample size); numbers that are very different in magnitude compared to 0.035 for B-level DIF in Jodoin and Gierl (2001). A suggestion is made in Cuevas and Cervantes (2012) that the Jodoin and Gierl (2001) cut-offs may be too conservative at identifying DIF, joining similar proposals put forward by Oliveri, Ercikan, and Zumbo (2014) and French and Maller (2007).

**Applications of Decision Rules in Simulation Studies**
Some of the decision rules have been tested in various simulation-based contexts, the results of which were primarily concerned with the Type I error and power rates. Jodoin and Gierl (2001) reported that Type I error had decreased when the effect size was included in the decision rule. The starkest note in regards to Type I error when using a blended decision rule, however, comes from French and Maller’s (2007) study, which concluded that with a blended decision rule, Type I errors decreased to “essentially zero” (p. 380). This result was particularly noticeable when unequal ability distributions were induced within the simulation. French and Maller (2007) noted that power was overly conservative, especially when the effect size was included into the decision rule, which echoed a similar statement made by Jodoin and Gierl (2001).

It is interesting to note that similar overall conclusions are drawn by both research teams, despite a difference in the decision rules that each study used. French and Maller’s (2007) results are explicitly only based on a one-degree of freedom chi-square test using an alpha of 0.01, not considering the possibility of the two-degree of freedom chi-square for the statistical significance criterion. On the other hand, Jodoin and Gierl (2001) considered both the one- and two-degree of freedom tests, but used an alpha of 0.05 under both conditions.

Hidalgo and López-Pina (2004) also performed a simulation study in comparing the statistical power of the M-H and the LR methods, with the LR methods also including the effect size. A large benefit of their approach is the inclusion of their Tables 2 and 3 (pp. 912-913), that highlight the mean and standard deviations of the effect sizes (specifically $\Delta R^2$ between the combined models). However, a large difference between the Hidalgo and López-Pina (2004) study, as compared to French and Maller (2007) and Jodoin and Gierl (2001), is where each one had applied the $\Delta R^2$ cut-offs. In contrast to Jodoin and Gierl (2001), Hidalgo and López-Pina
(2004) used the difference in $R^2$ between the first and the last LR models (i.e., the model including the interaction effect to represent non-uniform DIF) against the cut-offs suggested by Jodoin and Gierl (2001). The same approach was also used by Oliveri, Ercikan, and Zumbo (2014). The use of one, global, cut-off value was originally suggested by Zumbo (1999; albeit with the much more liberal cut-offs). That being said, Zumbo suggests that retaining the two sequential $\Delta R^2$ s would still be beneficial as the two values could be compared against each other to determine if the item is primarily uniform or non-uniform DIF. This approach may be more beneficial as it does not require the creation of separate $\Delta R^2$ cut-offs for both types of DIF, as is suggested in Jodoin and Gierl (2001). These results indicate that not only have there been a variety of cut-offs that have been suggested in the literature, but also differences in how they have been applied.

Statement of the Research Questions

As can be seen from this discussion, it is not only a question of what the cut-offs should be, but also a matter of clearly defining where the cut-offs are applied in the modelling process. It is important that in proposing the following research agenda that clearly articulated reasoning is used to outline each sequential step, allowing for further replication of the study method, with the hope of harmonizing the recommendations made to-date throughout the literature.

The first research goal to be considered here is a replication of an issue discussed in Jodoin and Gierl (2001): the conversion of the M-H cut-offs onto a scale used by the LR ($\Delta R^2$). As highlighted above, the recommendation presented in Jodoin and Gierl (2001) was developed with an intermediary conversion through SIBTEST and was based on datasets available to those researchers. An attempt will be made to replicate the existing methodology described in the literature with the use of a different dataset. Then, a more direct conversion with another dataset
will demonstrate if the cut-offs are interchangeable regardless of method by which they were calculated or which dataset was used.

Given new cut-offs that may lead from the first research question, the comparability of their results with the Jodoin and Gierl (2001) cut-offs, as well as with the original M-H cut-offs, will be examined. This will serve as a verification step to demonstrate the degree to which the three decision rules are interchangeable within each other. Since both of the LR decision rules were derived originally from the M-H, it is expected that they will be highly comparable.

Secondly, a broader question will be asked in regards to effect size: does their inclusion lead to different decisions at the item level, over and above the statistical test alone? In this case, how consistent are the decisions regarding DIF when the following decision rules are compared with each other:

1. LR: statistical significance – using $G^2$ statistical significance of the two-$df$ test only, with an alpha of 0.05 (based off Jodoin & Gierl, 2001)
2. LR: statistical significance – using $G^2$ statistical significance of either of the one-$df$ tests only, with an alpha of 0.05
3. LR: effect size – using Nagelkerke’s (1991) values to obtain $\Delta R^2$, applying cut-offs suggested by Jodoin & Gierl (2001) to the separate model changes (models 1 to 2, models 2 to 3)
4. LR: effect size – same as (3), with the exception of the application of the cut-off to the $\Delta R^2$ from model 1 through 3
5. LR: effect size – using $\Delta R^2$ based on Nagelkerke (1991) values, with any proposed cut-offs that may be developed from research question 1
6. LR: combined rule – rule (1) with rule (4)
7. LR: combined rule – rule (1) with rule (5)

Methods

In order to answer the questions above, the data source to be used for the study and the methods of the analysis will be detailed, in turn, below.

Data Source

The data source used for this analysis is drawn from an operational test used to measure functional English language use. Both the listening and reading components are each composed of 38 multiple-choice, single response, dichotomously-scored items. The majority of these items are grouped into testlets sharing a similar stimulus (e.g., a reading passage, or a recorded conversation). Despite the presence of testlets, the unidimensionality of all forms of this test has been examined and documented elsewhere. The choice of operational data for this study, rather than simulated data, will allow the techniques to be tested in an environment which includes the natural variability that is expected and observed in an operational setting.

In addition to operational item response data, the following demographic variables were used in an attempt to detect DIF:

- gender groups of the test taker, split into the two groups: “male”, and “female” (test takingers reporting other self-identified gender identities are excluded from the dataset)
- citizenship groups of the test taker, split into two groups: one group including those from a country primarily speaking English in a sociocultural context judged to be highly similar to that found in Canada (i.e., test takers who hold an initial citizenship of one of: Canada, United States, Australia, New Zealand, United Kingdom, or Ireland), and the other group consisting of those reporting other citizenships.

The analyses to follow will be symbolically coded with one (1) representing males and socioculturally similar groupings, and zero (0) representing females and test takers from
socioculturally dissimilar countries. Given the data splits explained above, the following table summarizes the number of test takers within each grouping, as well as the total size of the dataset:

Table 1. Sample size under various conditions for the four forms under consideration in this work.

<table>
<thead>
<tr>
<th>Form</th>
<th>N</th>
<th>males</th>
<th>females</th>
<th>similar context</th>
<th>dissimilar context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening A</td>
<td>2666</td>
<td>1674 (62.8%)</td>
<td>992 (37.2%)</td>
<td>254 (9.5%)</td>
<td>2412 (90.5%)</td>
</tr>
<tr>
<td>Listening B</td>
<td>2563</td>
<td>1597 (62.3%)</td>
<td>966 (37.7%)</td>
<td>236 (9.2%)</td>
<td>2327 (90.8%)</td>
</tr>
<tr>
<td>Reading A</td>
<td>2630</td>
<td>1859 (70.7%)</td>
<td>771 (29.3%)</td>
<td>363 (13.8%)</td>
<td>2267 (86.2%)</td>
</tr>
<tr>
<td>Reading B</td>
<td>2691</td>
<td>1866 (69.3%)</td>
<td>825 (30.7%)</td>
<td>397 (14.8%)</td>
<td>2294 (85.2%)</td>
</tr>
</tbody>
</table>

The purpose of the recommended splits is to demonstrate the comparability of the statistical techniques in two circumstances, especially in a circumstance that is more balanced than the other. Each of the previously mentioned studies related to LR have considered unbalanced group sizes. Since this is a common factor in those studies, it is also considered in this research context.

Although other research has usually used the simple total score as an estimate of ability, this work will use a corrected item-total: a total summed from the items taken by the test taker without taking into account the individual item being considered in a particular DIF regression. This prevents the total score from being conflated with the particular item should it, in fact, contain DIF.
**Question 1: Translating M-H Cut-Offs to LR**

First and foremost, the conversion of the M-H metric onto one used by LR was performed. This was accomplished using two methods. At first, an attempt to replicate Jodoin and Gierl’s (2001) results was performed. Second, since M-H has a metric that resembles a conditional odds ratio, this metric was transformed to the conditional odds ratio that was calculated from the coefficient estimated for the regression model 2 (the model which contains only ability and group membership as the predictors).

For each of the two data splits (gender and context similarity), the relationship between the M-H alpha scale, the conditional odds ratio, as well as the calculated $\Delta R^2$ for the transition from model 1 to 2 was investigated graphically. The goal of graphical relationship modelling was to investigate the whether or not the various datasets could be merged into one translating model (cf. Jodoin and Gierl, 2001). From the fitted model(s), the M-H delta cut-offs were transformed onto the $\Delta R^2$ metric.

The most notable drawback of this approach is that it only explicitly uses the regression coefficient for group membership from model 2, which, in fact, only captures the condition of uniform DIF. However, one could make an argument that the measure of DIF should not depend on the type of DIF. With this reasoning, if a group difference is considered large enough by uniform standards when using a strategy that is only designed to detect uniform differences in items, then differences that are just as large with partially (or exclusively) non-uniform DIF should also be considered as equally suspect. As a result, the recommendation provided in this study will be that the cut-off, even though it is developed exclusively from a source describing uniform DIF, can (and should) be applied to the difference in $R^2$ from models 1 to 3. In other words, the degree of DIF, as measured by effect size (a practical difference), should not be
dependent on the type of DIF (a theoretical description of the type of DIF). With this strategy, the recommendation made by Zumbo (1999) can be used: the ratio of the two different $R^2$ changes can be compared to identify whether or not the item exhibits primarily uniform or non-uniform DIF. For example, if 80% of the item’s $R^2$ difference occurs in the transition between models 2 and 3, one can reasonably conclude that the item exhibits primarily non-uniform DIF. From a strictly pragmatic orientation, the concern surrounds the amount of DIF (the effect size), as opposed to how that DIF is manifested (i.e., uniform vs. non-uniform).

**Question 2: Comparability of Results**

Next, the various techniques will be run on the datasets described earlier. These techniques have been programmed into a computer script written in R (R Core Development Team, 2014). This automated process will output the raw results of the various statistical tests outlined earlier, as well as the interpretations suggested by the various decision rules outlined earlier.

To address the comparability between the different measures, a consistency matrix was developed. This matrix will show percent agreement on the off-diagonal, and the proportion of items classified at each level on the diagonal. Since the decision rules result in either two or three categories of DIF classification (either the “yes”/“no” of statistical significance, or the A-, B-, and C-levels of the combined decision rule), B- and C-level items will be consolidated into one category for the purposes of calculating agreement, representing a presence/absence of DIF.
Results

Results for Question 1: Translating M-H Cut-Offs to LR

Distributions of test taker ability. At first, the comparison of the raw ability distributions are demonstrated between the two types of data groupings. From the following graphs, it is possible to see that gender grouping represents a group of test takers who are highly overlapped (see Figure 3), compared with the highly non-overlapping nature of test takers split by sociocultural context (see Figure 4). This has implications for DIF (Zumbo, 1999), as matching becomes a statistical creation. As a result, although the sociocultural difference may be a DIF result faced in actual data (as in this case), the majority of calculations in this section will be based on the gender-based DIF results. Results in Figure 3 and Figure 4 present variables coded as one (1) in solid lines, and zero (0) in dashed lines.
Figure 3. Density plot of summed score by gender, in both listening forms (left) and reading forms (right).
Method 1: Attempting Replication of Jodoin & Gierl (2001). At first, an attempt was made to replicate Jodoin and Gierl’s (2001) methodology in the determination of cut-offs. SIBTEST was run in DIF-Pack version 1.7 (William Stout Institute for Measurement, 2005). In keeping with the method in Roussos and Stout (1996), as cited by Jodoin and Gierl (2001), each item was considered for DIF against all other items on the test. As well, a bi-directional test of DIF was used (no directionality of differential functioning was assumed in any of the datasets). With the intention of replication, only the gender-based DIF was considered for this sub-study.

The results are presented in Figure 5, below. Two solid vertical lines were placed on the graph to represent the two locations of the SIBTEST cut-offs (0.059 and 0.088 on the $|\hat{\beta}|$ axis.

Figure 4. Density plot of summed score by sociocultural background, in both listening forms (left) and reading forms (right).
The dashed line superimposed on the data is a cubic regression, resembling the functional form suggested by Jodoin and Gierl (2001). To contrast, a solid line on the dataset shows the loess-based representation of the data.

![Comparing SIBTEST vs. LR](image)

**Figure 5.** SIBTEST $|\hat{\beta}|$ on x-axis with LR-based $\Delta R^2$ between models 1 and 2 on the y-axis.

Based on the loess curve, the cubic function may not necessarily resemble the data the best in the lower range (especially around SIBTEST $|\hat{\beta}| < 0.025$). The functional form is a better representation of the loess mapping in the range where the SIBTEST-based cut-offs exist; however, there is a notable lack of data as well in the same range. For consistency in method to the original article, the cubic functional form best-fit to this data is calculated as:

$$\Delta \hat{R}^2 = 14.93691 \beta^3 - 0.79805 \beta^2 + 0.05732 \beta + 0.00018,$$

leading to suggested cut-offs of 0.00385 and 0.00922 for B-level and C-level DIF.
These results diverge strongly from the recommendations provided in Jodoin and Gierl (2001) of 0.035 and 0.070. To emphasize the divergence of results, Figure 5 from above is included again, rescaled to match the ranges of Figure 1 from Jodoin and Gierl (2001, p. 336). Upon comparison of the figures, the following two differences are apparent: (a) the range of $\Delta R^2$ in the original work is much wider than that seen with this data set, with $\Delta R^2$ values as high as ~0.65 reported compared to a maximum of 0.0095 shown here; and, (b) shape of the cubic regression model is visibly different. The large outliers in the items analyzed in Jodoin and Gierl (2001) may also have highly influenced the nature of the cubic regression model because of their high leverage values.

Figure 6. Rescaling of Figure 5 to match the presentation of Figure 1 from Jodoin and Gierl (2001, p. 336).
Furthermore, the linear relationship between SIBTEST $\hat{\beta}$ and M-H's $\hat{\lambda}$ was verified visually (Roussos & Stout, 1996). However, the obtained correlation of 0.44 found with this data distinctly differed from a previously reported correlation of 0.98 (Dorans, 1991), which was used as the basis of the transformation between the M-H and SIBTEST DIF detection techniques (Roussos & Stout, 1996).

Because the results using SIBTEST are divergent in both (a) M-H to SIBTEST, and (b) SIBTEST to LR, a direct transformation of M-H to LR may be a simpler and more defensible strategy.

**Method 2: Directly Linking M-H and LR.** As a result of these large differences, and in an attempt to simplify the transition between M-H and LR cut-offs, another strategy was considered in mirroring these two DIF detection techniques. This technique, as discussed earlier, directly maps the conditional odds ratios between both techniques, then considers their relationship with the change in $R^2$.

**Checking the expected relationship between delta (transformed logOdds in M-H) and logOdds from LR.** Firstly, the conditional odds of group membership from both the M-H and the LR methods are compared to demonstrate that the two are indeed interchangeable. For the purpose of ease of interpretation, the following graph demonstrates the linearized versions of both. The logOdds was multiplied by 2.35 to match the scaling of the delta metric, and is demonstrated below with linear regressions for each DIF type.
Figure 7. Relationship between deltaMH and the logOdds of Group in model 2, placed on the same metric scale, with superimposed linear regressions. Points representing gender DIF are clustered towards the bottom-right of the graph and are represented with circles, with socioculturally data clustered to the top-left of gender and represented with triangles.

The relationship between the two variables for gender estimated by linear regression is 
\[ \hat{y} = -0.04252 + 0.97649 \, x \]. Socioculturally similar groupings have a relationship between the two variables of \[ \hat{y} = 0.85750 + 0.93770 \, x \]. The graph and data both reveal that the relationship (slope) between the two variables, in both regressions, tends towards 1; however, only the gender-based DIF reveals the expected identity relationship with an intercept of approximately 0 (Holland & Thayer, 1988). Because the sociocultural similar grouping does not match the expected relationship between the two variables, the gender-based models will be used in the remainder of this work to construct a new recommended cut on the \( \Delta R^2 \) scale. The unexpected performance of the data from sociocultural background DIF is likely related to the non-overlapping ability distributions between the two groupings. With the interchangeability of the two scales verified
for the gender-based divisions, a $|\Delta_{\text{MH}}| = 1$ would, after dividing by 2.35 (the scaling factor), equal a logOdds of approximately 0.42553, along with a $|\Delta_{\text{MH}}| = 1.5$ equalling a logOdds of 0.63830.

**Converting logOdds to $\Delta R^2$.** It is necessary to convert the logOdds of group membership in model 2 to the $\Delta R^2$ metric, in order to enable DIF determinations to occur in the broader and more encompassing model 3 (the model which includes non-uniform DIF). $\Delta R^2$ remains the most convenient effect size measure for LR as the difference in $R^2$ between models 1 through 2 is nested within the difference in $R^2$ between model 1 and the non-uniform accounting model 3. Since both the (conditional) odds ratio and $\Delta R^2$ are measures of effect size, it is appropriate to relate the two variables; which is visualized below. Given the conclusion from above that the gender DIF more accurately respects the relationship expected between the conditional odds ratio of M-H and the conditional odds ratio of LR, only its results are used for the purposes of conversion. Two methods; an empirically-derived and a model-based approach are considered next.

To better reflect the actual data, an empirical-based strategy was first considered. In this strategy, increasing ranges around the target logOdds (0.42553 and 0.63830) are built; within each range, the median is calculated. In the model-based strategy, an estimated mathematical model was selected that best resembled the loess line suggested by the data.

**Results of the empirical strategy.** The empirical strategy was initially preferred over a model-based strategy, as it did not impose a functional form on two variables whose relationship has not clearly been linked. The results of the empirical-based strategy are shown below, in Figure 8, for the target logOdds of 0.42553. In this strategy, sequential steps in increments of
0.0005 above and below the target logOdds were taken for a duration of 100 steps (i.e., a total range of 0.42553 ± 0.05). At each step, the median was calculated, and is presented on the y-axis. The figure only begins to display data when the step contained more than two data points (which occurred starting at step 39). With the exception of a small dip between steps 78 and 81, the median is stable at 0.00686 until the end, calculated on the basis of five data points. The dotted line represents the model-based result for the same cut point (detailed later), showcasing the similarity of results between the two strategies.

Figure 8. A graph demonstrating the median results of the increasing window around the target logOdds of 0.42553. The dotted horizontal line demonstrates the model-based estimate, discussed later in text.

An attempt was made to compute the median ΔR² at the cut-point of 0.63830; however, there was no data available for gender-based DIF at this level. In order to be able to calculate an
equivalent in this range, a model-based imputation – although not preferred – is a required alternative.

Results of the model-based strategy. The functional form that best resembled the data is an exponential model. For the purposes of modelling, the graph below also shows a loess curve (solid line), and a modelled power relationship (dotted line). An acknowledgement is made that the functional form selected for the model is strictly based on best-fit to the data, and not on any proofs relating the two variables. As well, because data is not available in the range of the second cut-off, the model is only a best-informed guess of what possible values could be should there have been data in that range.

Figure 9. The relationship between the absolute value of the logOdds of group in model 2 as opposed to $\Delta R^2$ between models 1 and 2. The loess curves are plotted using solid lines; matching modelled relationships (power) in dotted lines.
In this case, the relationship between the logOdds and $\Delta R^2$ between model 1 and 2, for gender-based DIF, is estimated by the formula: $\Delta \hat{R}^2 = 0.03665 \log \text{Odds}^{1.96970}$. Given this estimated relationship, a cut-off of 1 on $|\Delta_{\text{hit}}|$ solves to a $\Delta \hat{R}^2$ of 0.00681, and 1.5 solves to a $\Delta \hat{R}^2$ of 0.01514.

The model-estimated cut-off for the first cut point was graphed as a horizontal line in Figure 8. This shows concordance between the mean-based regression technique, and the data-based median. Since the empirical technique does not provide results for the second cut, it is suggested that both cut-offs be model-based to preserve consistency, and because of relatively close agreement of the first cut with the empirical method.

**Results for Question 2: Comparability of Results**

Having suggested different cut-off values with the use of a different methodology, the cross-comparability of the various methods are considered in turn. At first, a general summary of the comparability of the various methods is presented in the following two tables. Table 2 demonstrates the application of the cut-offs for all data points (both types of DIF under consideration in this study). Table 3 demonstrates the same for the gender-only DIF situation (e.g., the context within which equal ability distributions are seen). In both tables, the diagonals show the results of classification for that particular decision rule; the off-diagonals show the percentage of exact agreement between two decision rules. In cases where a decision rule consisting solely of a significance test is compared against a decision rule generating A/B/C categories, B- and C-level items were collapsed into one category of items which exhibit DIF (similar strategy used in Jodoin & Gierl, 2001).
Table 2. A table of exact agreement (off-diagonal) and proportions (diagonal) of various decision rules applied to all types of DIF under consideration.

<table>
<thead>
<tr>
<th></th>
<th>ALL DATA</th>
<th>Statistical Significance Only</th>
<th>Effect Size Only</th>
<th>Blended Decision Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chi-Square 2df</td>
<td>Non-Significant = 51% Significant = 49%</td>
<td>.93</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>Chi-Square 1df, either model</td>
<td>TRUE = 46% FALSE = 54%</td>
<td>.46</td>
<td>.48</td>
</tr>
<tr>
<td></td>
<td>alpha = 0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 2 or 2 to 3 (Jodoin &amp; Gierl, 2001)</td>
<td>.51</td>
<td>.46</td>
<td>A = 100%</td>
</tr>
<tr>
<td></td>
<td>(Jodoin &amp; Gierl, 2001)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 3 (Jodoin &amp; Gierl, 2001)</td>
<td>.53</td>
<td>.48</td>
<td>A = 98%  B = 2%</td>
</tr>
<tr>
<td></td>
<td>(new cuts)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 3 (new cuts)</td>
<td>.73</td>
<td>.68</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>(Jodoin &amp; Gierl, 2001)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 3 (Jodoin &amp; Gierl, 2001) with Chi-Square</td>
<td>.53</td>
<td>.48</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td>2df &lt; 0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 3 (new cuts) with Chi-Square 2df &lt; 0.05</td>
<td>.73</td>
<td>.68</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>(Holland &amp; Thayer, 1988)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 3 (new cuts) with Chi-Square</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2df &lt; 0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 3 (Jodoin &amp; Gierl, 2001) with Chi-Square</td>
<td>.53</td>
<td>.48</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td>2df &lt; 0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 3 (new cuts) with Chi-Square 2df &lt; 0.05</td>
<td>.73</td>
<td>.68</td>
<td>.78</td>
</tr>
<tr>
<td></td>
<td>(Holland &amp; Thayer, 1988)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Change in R^2 in Models 1 to 3 (new cuts) with Chi-Square</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2df &lt; 0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3. A table of exact agreement (off-diagonal) and proportions (diagonal) of various decision rules applied to gender DIF data only

<table>
<thead>
<tr>
<th>GENDER DIF ONLY</th>
<th>Statistical Significance Only</th>
<th>Effect Size Only</th>
<th>Blended Decision Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chi-Square 2df, alpha = 0.05</td>
<td>Chi-Square 1df, either model alpha = 0.05</td>
<td>Change in R^2 in Models 1 to 2 or 2 to 3 (Jodoin &amp; Gierl, 2001)</td>
</tr>
<tr>
<td>Statistical Significance Only</td>
<td>Non-Significant = 69% Significant = 31%</td>
<td>.90</td>
<td>.69</td>
</tr>
<tr>
<td>Effect Size Only</td>
<td>Change in R^2 in Models 1 to 2 or 2 to 3 (Jodoin &amp; Gierl, 2001)</td>
<td>.90</td>
<td>.62</td>
</tr>
<tr>
<td>Change in R^2 in Models 1 to 3 (Jodoin &amp; Gierl, 2001)</td>
<td>.69</td>
<td>.62</td>
<td>.69</td>
</tr>
<tr>
<td>Change in R^2 in Models 1 to 3 (new cuts)</td>
<td>.75</td>
<td>.68</td>
<td>.94</td>
</tr>
<tr>
<td>Blended Decision Rules</td>
<td>Change in R^2 in Models 1 to 3 (Jodoin &amp; Gierl, 2001) with Chi-Square 2df &lt; 0.05</td>
<td>.69</td>
<td>.62</td>
</tr>
<tr>
<td>Change in R^2 in Models 1 to 3 (new cuts) with Chi-Square 2df &lt; 0.05</td>
<td>.75</td>
<td>.68</td>
<td>.94</td>
</tr>
<tr>
<td>(Holland &amp; Thayer, 1988)</td>
<td>.72</td>
<td>.64</td>
<td>.97</td>
</tr>
</tbody>
</table>
The tables above reveal that very few items are classified as exhibiting DIF when the cut-offs recommended by Jodoin and Gierl (2001) are applied. The addition of effect size clearly reduces the number of items identified as DIF, as highlighted in the literature. However, in the data used for this study, the blending of statistical significance and effect sizes yields nothing additional when compared to effect sizes used alone. That noted, the retention of the statistical significance is required to keep with the tenets of hypothesis testing. Although exact agreement is very similar between M-H (Holland & Thayer, 1988), Jodoin and Gierl (2001), and the cut-offs presented here; the distribution of A-/B-/C-level identified items are much more similar between the cut-offs presented here than those presented by Jodoin and Gierl (2001).

Despite the similarity in proportions, it is worthwhile to look closer at the distribution of items as identified by the blended decision rule proposed here against the decision rules as applied to the M-H technique (see Table 4).
Table 4. Classification contingency matrix for DIF items using decision rules from Holland & Thayer (1988) for M-H vs. those calculated in this work, all numbers as proportions of total item mix for all data.

This table demonstrates that 78% of items (i.e., the sum of the diagonal) were classified in the same manner by both techniques. The items present in the categories of most disagreement (i.e., items identified “A” with one technique and “C” with the other) occurred within items tested for sociodemographic DIF, which comes with the challenge of an unequal ability distribution demonstrated earlier. Ten percent of items were classified higher using the decision rule recommended in this work, than the decision rules used with the M-H technique. The higher classifications using decision rules from this work may be related to the presence of non-uniform DIF. Figure 10, below, shows that items classified higher (bottom of graph) do tend to have
higher $\Delta R^2$ s between models 2 and 3. Naturally, this is to be expected as the M-H is not intended to test for non-uniform DIF. No obvious relationship was observed to explain why items may be classified lower using the decision rule proposed here when compared to M-H.

Figure 10. A graph comparing the distribution of $\Delta R^2$ s between Models 2 and 3. The top graph shows the distribution of $\Delta R^2$ for items classified equal / lower using decision rules suggested here, when compared to those suggested by Holland & Thayer. The bottom graph shows the distribution of $\Delta R^2$ for items classified higher (containing more DIF) using the new decision rule as opposed to their Holland & Thayer classification.
Concluding Remarks

In effect, despite the known ability to use the LR technique for DIF detection over the last 15 years, there has yet to be a single decision rule used consistently within the literature. The intention of this research was to identify a comparable cut-off in the LR test from the M-H, and to perform a comparison between the various DIF flagging techniques presented in the literature so far.

The work here revealed suggested cut-offs on the $\Delta R^2$ metric that are inconsistent with findings in the current literature, summarized in the table below (Table 5). Throughout the work, substantial differences were found when unequal ability distributions were considered, prompting the need for further evaluation of the LR technique under these specific conditions. There are indications that LR is more sensitive than M-H to non-uniform DIF, as defined by the change in $R^2$ between LR models 2 and 3, supporting the idea that LR can capture non-uniform DIF.
Table 5. Summary of differences in cut-offs on the $\Delta R^2$ metric between various sources used to-date in the literature.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>models 1 to 2 or models 2 to 3</td>
<td>models 1 to 2</td>
<td>models 2 to 3</td>
</tr>
<tr>
<td>A-level</td>
<td>$\Delta R^2 &lt; 0.035$</td>
<td>$\Delta R^2 &lt; 0.000363$</td>
<td>$\Delta R^2 &lt; 0.00062$</td>
</tr>
<tr>
<td>B-level</td>
<td>$0.035 &lt; \Delta R^2 &lt; 0.070$</td>
<td>$0.000363 &lt; \Delta R^2 &lt; 0.001012$</td>
<td>$0.00062 &lt; \Delta R^2 &lt; 0.002708$</td>
</tr>
<tr>
<td>C-level</td>
<td>$0.070 &lt; \Delta R^2$</td>
<td>$0.001012 &lt; \Delta R^2$</td>
<td>$0.002708 &lt; \Delta R^2$</td>
</tr>
</tbody>
</table>

Note. The number of digits after the decimal point are presented as reported in the original articles in order to preserve the authenticity of the recommended cut-offs.

$^a$The units of the Cuevas and Cervantes (2012) are on the metric of McFadden’s $\Delta R^2$, and therefore, are not directly comparable with the remaining values in the table. As well, due to the multitude of cut-offs presented in the original work, the largest cut-offs in each series were selected to present in this summary table. The selected cut-offs also resemble the gender condition presented here the closest, in terms of sample size and proportion between the two groups.
On reviewing the table, it is apparent that the scale of the cut-off values on the $\Delta R^2$ metric are notably miniscule. This is especially the case when the cut-off values are compared with the range $R^2$ s are able to take (0 to 1), as well as the effect size guidelines suggested by Cohen (1992). This may suggest to a researcher, unaccustomed to these values as applied in the context of DIF, to conclude that such results that are inconsequential. Prentice and Miller (1992) specifically discuss how numerically small effect sizes may still reflect an important change worth measuring, such as the case where effects are measured when the dependent variable is “unlikely to yield to influence from the independent variable” (p. 162). Such broad examples of groupings that are customary within the DIF literature, such as “male” or “female”, combined with a large number of test-takers within each grouping, would reasonably point to a more heterogeneous group than a homogeneous group. As a result, even if a tiny difference is found in item score between two gender groups – which consist of a wide variety of individuals of different ages, socioeconomic backgrounds, and educational experiences to name a few – may demonstrate that gender can operate “even in domains [one] would think were immune to its effects” (Prentice & Miller, 1992, p. 162). Therefore, such small changes in $R^2$ are still worth monitoring for bias.

**Implications for Practice**

As can be seen in this work, a new cut-off is suggested that is notably dissimilar from that suggested by Jodoin and Gierl (2001). This may be due to differences in the technique in which the results were calculated, or differences in the underlying dataset. Although the results used in this study come from two different constructs (listening and reading), both scales originate from the same assessment instrument and the same population. Consequently, the cut-offs that are
calculated as a result of this research may actually be highly tied to the dataset used, including the nature of the construct or changes in the distribution of the item parameters.

It is clear that the need for effect sizes within DIF are not only necessary, but have been used for a number of years. Nonetheless, the selection of effect sizes for screening for DIF must be done with purpose. Psychometricians and other researchers must always remember that the effect sizes for DIF, like all statistical effect sizes, must always be interpreted in the context in which they are used. The idea of establishing one universal effect size that can be used without contextualization is a high order; Cohen specifically warned against universal effect sizes (1988). Despite this, DIF techniques seem to assume the universality of their cut-offs. In some cases, such as an initial exploratory study of a construct, a higher level of DIF may be permissible than traditionally used. In other cases, psychometricians may be required to use a set of more constrained cut-offs, such as for items on a high-stakes limited-number-of-attempts assessment product. These categories based on research questions may still be too broad: Do different constructs require different cut-offs? Should different test taker groupings be allowed to have different thresholds for the identification of DIF?

All in all, effect sizes selected must be defensible, respecting test taker groupings that may be affected by an item exhibiting DIF. It is worth considering, with the goal of increasing the defensibility of a testing programme’s DIF cut-offs, that one could draw upon procedures traditionally used in standard setting for cut score development (cf. Cizek, 2011). If this DIF cut-off setting procedure is carried out, including enough varied panelists, with test-, construct-, and test use-specific knowledge, the defensibility of cut-offs should be higher than the application of universally established cut-offs. Testing programmes should then consider the collection of
validity evidence to support their DIF cut-offs, much like cut scores require validity argumentation.

As a result, it is recommended that psychometric practitioners evaluate potential cut-offs for DIF on a testing programme-specific basis. The difference in results presented here as opposed to those presented earlier in the literature suggests that different DIF cut-offs may occur as an interaction of test items, test domain, and testing population. Should any of the three criteria change, the cut-off used to evaluate DIF items may also require adjusting. The use of a single universal cut is, therefore, not recommended.

**Other Directions for Research**

An acknowledgement is made that the proposed methods are only a stepping stone for future research in this domain. First and foremost, this study cannot clearly be considered a power, Type I, or a Type II study of the various statistical techniques. Without a simulation study, one cannot determine the actual detection rate of DIF as one has not artificially inserted performance differences into the data. It would be relevant to follow up this study with simulation studies to further investigate the ways that various DIF decision rules perform, including the proposed cut-offs presented in this work.

As well, because this data is taken from an operational test, the number of items identified as containing DIF were limited. With limited numbers of items identified as DIF, cross-comparability between the various measures is also limited. Repeating this experiment with purposefully simulated data, with specifically inserted DIF, may increase the ability to draw any conclusions about the cross-comparability of measures.
References


Type (Ordinal) Item Scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223–233.

